

On Optimal Top- K String Retrieval

Rahul Shah¹, Cheng Sheng², Sharma V. Thankachan¹, and Jeffrey Scott Vitter³

¹ Louisiana State University, USA. {rahul, thanks}@csc.lsu.edu

² The Chinese University of Hong Kong, China. csheng@cse.cuhk.edu.hk

³ The University of Kansas, USA. jsv@ku.edu

Abstract. Let $\mathcal{D} = \{d_1, d_2, d_3, \dots, d_D\}$ be a given set of D documents of total length N . The top- K document retrieval problem is to index \mathcal{D} such that when a pattern P and a parameter K comes as a query, the index returns the K most relevant documents to the pattern P (in sorted order). Hon et al. [15] gave the first linear space framework to solve this problem in $O(|P| + K \log K)$ time. This was improved by Navarro and Nekrich [22] to $O(|P| + K)$. In these papers, first a pattern matching is done and then the documents are retrieved based on the locus of the pattern. During the retrieval phase, the factor $O(P)$ is used to bound the query time. Once one considers word-packing in RAM or external memory model, this factor is no longer optimal. Besides many applications require retrieval to be independent of pattern searching. We show a linear index which takes strictly $O(K)$ time, once the locus of pattern match is given. Separately, we also give an external memory linear space index taking near-optimal $O(|P|/B + \log_B N + \log \log B + K/B)$ I/Os (outputs not sorted). Our results are surprising in the sense that they defy the usual range searching bounds. Our techniques also have implications in cache-oblivious model.

1 Introduction

The inverted index is the most fundamental data structure in the field of information retrieval [30]. It is the backbone of every known search engine today. For each word in any document collection, the inverted index maintains a list of all documents in that collection which contain the word. Despite its power to answer various types of queries, the inverted index becomes inefficient, for example, when queries are phrases instead of words. This inefficiency comes from inadequate use of word orderings in query phrases [24]. Similar problems also occur in applications when word boundaries do not exist or cannot be identified deterministically in the documents, like genome sequences in bioinformatics and text in many East-Asian languages. These applications call for data structures to answer queries in a more general form, that is, (string) pattern matching. Specifically, they demand the ability to identify efficiently all the documents that contain a specific pattern as a substring. The usual inverted-index approach might require the maintenance of document lists for all possible substrings of the documents. This can take quadratic space and hence is neither theoretically interesting nor sensible from a practical viewpoint.

The first frameworks for answering pattern matching (and related) queries were proposed by Matias et. al. [20] and Muthukrishnan [21]. Their data structures solve the *document listing problem*, in which a collection \mathcal{D} of D documents is required to be indexed so that given a query pattern P , all the documents that contain P can be retrieved efficiently. As the pattern can appear in a single document multiple times, a major challenge of this problem is that the overall number of the pattern occurrences can be much greater than the number $ndoc$ of the result documents. Therefore, it is unaffordable to answer a query by enumerating all the occurrences of P .

Muthukrishnan also initiated the study of relevance metric-based document retrieval [21], which was then formalized by Hon et al. [15] as the *top- K document retrieval problem*. Here, instead of all the documents that match a query pattern, the problem is to output the K documents most relevant to the query in sorted order of relevance score. Relevance metrics considered in the problem can be either pattern-independent (eg., PageRank) or -dependent. In the latter case one can take into account information like the frequency of the pattern occurrences (or term-frequency of popular tf-idf measure) and even the locations of the occurrences (e.g., *min-dist* [15] which takes proximity of two closest occurrences of pattern as the score). The framework of Hon et al. [15] takes linear space and answers the query in $O(|P| + K \log K)$ time. This was then improved by Navarro and Nekrich [22] to achieve $O(|P| + K)$ query cost, which is in a way optimal. Several other approaches for top- K document retrieval have recently been published. Some use, instead of linear space, succinct space [5,15] or semi-succinct space [29,11,23,14,5]. Their query costs, however, usually contain a multiplicative poly-logarithmic factor to the output size K (or $ndoc$). This problem is seeing a burst of research activity in both mainstream venues for algorithms and information retrieval [15,17,22,24] as well as plenary talks [23,16] in the string matching community.

All the above approaches use a two-phase procedure to answer a query. The first phase identifies the *locus* of P in a suffix tree, that is, the node corresponding to the pattern P . The second phase finds the top- K results in the subtree rooted at the locus. [15] and [22] reduce the Phase-2 subproblem to a 3d orthogonal range searching problem with four constraints. While general four-constraint orthogonal range searching is proved hard [9], the desired bounds can nevertheless be achieved by identifying a special property that one dimension of the reduced subproblem can only have $|P|$ distinct values. Employing this property, an additive $O(|P|)$ -term inevitably appears in the cost to handle Phase-2, which is actually sub-optimal. In this paper, we shall prove that Phase-2 can be answered strictly in $O(K)$ time in the RAM model. Our techniques also have implications in the external memory (EM) model. Following are some motivating applications:

1. None of the existing approaches work in external memory. Here, we need the $|P|/B$ factor in I/O to be optimal. Thus, it is expensive to have additive $O(|P|)$ factor. Even in RAM, the optimal pattern matching time (based on word-packing) is taken as $O(|P|/\log_{|\Sigma|} N + \log \log N)$. In this sense, $O(|P| + K)$ bound is not optimal even in RAM.

2. In applications like cross-document pattern matching [18], pattern P is given by a location in some document and is needed to be found in other documents. Since the collection can be pre-indexed, the locus of the pattern can be found in $O(\log \log N)$ using weighted level ancestor query (or even in faster $O(\log \log |P|)$ time). Thus, an additive $O(|P|)$ factor can be unaffordably expensive.
3. Autocompletion has become an indispensable component of modern search engines. For example, in Google InstantTM ⁴, instead of waiting for a user to complete a query before the search starts, relevant results will be rendered in real time as the user types. In the view of a server, if the user types a string P , this procedure can issue up to $|P|$ queries, one for each prefix of P . Therefore, answering a Phase 2 query in $O(|P| + K)$ time leads to an $O(|P|^2)$ term in the overall query cost.
4. In many pattern matching applications, for example in suffix-prefix overlap [28] or maximal substring matches [19], multiple loci are searched with amortized constant time for each locus. In such situations, having extra $O(|P|)$ from the retrieval part leads to non-optimal solutions.

1.1 Related Work, Problem Complexity and Our Approach

For the document listing problem, Muthukrishnan gave a somewhat optimal solution, which uses linear space and $O(|P| + ndoc)$ query cost [21]. As the overall number of the occurrences of P can be much larger than $ndoc$, he uses the idea of *chaining*, by which a one-sided constraint on a particular dimension guarantees that no document can be enumerated more than once. With proper labeling, the pattern can be converted into a 2d orthogonal three-sided query. Hon et al. extend the idea to tree-shaped chaining, which can be used to solve the top- K document retrieval problem [15]. Here, the extra top- K constraint can be converted to a threshold in the third dimension, thus making the query four-constraint query in 3d space.

Both [15] and [22] use the fact that on one of the dimensions, the set of the geometric points related to query P have only $|P|$ distinct values. There are some other properties that potentially also ease the problem, compared to the general orthogonal range searching problem. One of them is that two of the constraints form a *tree range*, that is, the range of the pre-order ranks of the subtree rooted at a node in the tree. This limits the number of possible ranges to $O(N)$ instead of $O(N^2)$. Chien et. al. [10] show, however, that any general range can be broken into a logarithmic number of tree ranges, which implies that the advantage of this speciality can improve only an additive poly-logarithmic term. Another speciality is that the third constraint always has the same boundary as the first constraint. By illustrating the first two constraints on the x-dimension and the third constraint on the y-dimension, the query projects to a *hinged* three-sided window, that is, one of its corner lies on the diagonal $y = x$. The top- K constraint translates to the fourth constraint on the third dimension. This is what is exploited in our external memory solution, which is near optimal. We further combine the property of hinged windows with tree ranges to show that these four-constraint queries can be broken into $O(\log N)$ three-sided subqueries, which leads to novel results in the cache-oblivious model.

1.2 Our Results

In RAM, Phase-1 (i.e., finding the locus of pattern P), can be processed in either $O(|P|)$ time with a suffix tree, or in $O(|P|/\log_{|\Sigma|} N + \log \log N)$ time with a word-packed suffix tree. In the external memory or cache-oblivious model, Phase-1 can be answered in $O(|P|/B + \log_B N)$ I/Os with a string B-tree [7]. We summarize our results as follows.

1. In RAM, there exists an $O(N)$ -word data structure that solves Phase-2 of the top- K document retrieval problem in optimally $O(K)$ time. This result improves the previous work [15,22] by eliminating the additive term $|P|$.
2. In the external memory model, there exists an $O(N)$ -word structure that solves the top- K document retrieval problem in $O(\log_B N + \log \log B + K/B)$ I/Os. This is surprising because the bound is closer to the three-constraint orthogonal range searching problem, instead of the four-constraint one (as

⁴ <http://www.google.com/insidesearch/features/instant/about.html>

the problem appears to be). Further the optimal $O(\log_B N + K/B)$ query I/Os can be achieved using an almost-linear $O(N \log \log B)$ -word space structure.

3. In the cache-oblivious model, there exists an $O(N)$ -word structure that solves the document listing problem in $O(\log N + ndoc/B)$ I/Os. Notice that this problem is at least as hard as the *interval stabbing problem*, and at most as hard as the three-sided orthogonal range searching problem. While $O(\log N)$ term does not look optimal, no better result is known for the interval stabbing problem in this model. On the other hand, it has been proved that any $O(\log^{O(1)} N + ndoc/B)$ -I/O structure must take super-linear $\Omega(N \log^\epsilon N)$ space [2]. This shows that the document listing problem is, by hardness, closer to the interval stabbing problem than to the three-sided orthogonal range searching problem. For top- K document retrieval problem, also one can derive new bounds nearly similar to 3-sided query bounds (using super-linear space) in the cache-oblivious model.

2 Top- K String Retrieval Framework

This section briefly explains the framework of Hon et. al. [15]. Define $score(P, d)$, the *score* of a document d with respect to a pattern P , to be the relevance of d to P , which is a function of the locations of all P 's occurrences in d . The generalized suffix tree (GST) of a document collection $\mathcal{D} = \{d_1, d_2, d_3, \dots, d_D\}$ is the combined compact trie (a.k.a. Patricia tree) of all the nonempty suffixes of all the documents. Use N to denote the total length of all the documents, which is also the number of the leaves in GST. For each node u in GST, consider the path from the root node to u . Let $depth(u)$ be the number of nodes on the path, and $prefix(u)$ be the string obtained by concatenating all the edge labels of the path. For a pattern P that appears in at least one document, the *locus* of P , denoted as u_P , is the node closest to the root satisfying that P is a prefix of $prefix(u_P)$. By numbering all the nodes in GST in the pre-order traversal manner, the part of GST relevant to P (i.e., the subtree rooted at u_P) can be represented as a range, called the *suffix range* of P .

Nodes are marked with documents. A leaf node ℓ is marked with a document $d \in \mathcal{D}$ if the suffix represented by ℓ belongs to d . An internal node u is marked with d if it is the lowest common ancestor of two leaves marked with d . Notice that a node can be marked with multiple documents. For each node u and each of its marked document d , define a *link* to be a quadruple $(origin, target, doc, score)$, where $origin = u$, $target$ is the lowest proper ancestor⁵ of u marked with d , $doc = d$ and $score = score(prefix(u), d)$. Two crucial properties have been identified in [15].

Lemma 1 *The total number of links is upper bounded by $O(N)$.*

Lemma 2 *For each document d that contains a pattern P , there is a unique link whose origin is in the subtree of u_P and whose target is a proper ancestor of u_P . The score of the link is exactly the score of d with respect to P .*

The top- K document retrieval problem can be thus reduced to the problem of finding the top- K links that originate in the subtree of u_P and target at a proper ancestor of u_P . With the nodes in GST numbered in the *pre-order traversal* order, these constraints translate into finding all the links (i) the numbers of whose origins fall in the number range of the subtree of u_P , and (ii) the numbers of whose targets are less than the number of u_P . Regarding constraint (i) as a two-sided range constraint on x-dimension, and constraint (ii) as a one-sided range constraint on y-dimension, the problem asks for the top- K points that fall in a three-sided window in 2d space. Furthermore, as the left endpoint of the range in (i) always equals the endpoint of the range in (ii), one corner of the three-sided window must be on the diagonal $y = x$. We thus name the resulting problem as *top- K hinged range reporting*.

3 Linear Space, $O(\log N + K)$ Retrieval Time Data Structure

Once the locus u_P of a pattern P has been identified, Hon et. al.'s structure retrieves the top- K documents in $O(|P| + K \log K)$ time [15], while Navarro and Nekrich's take $O(|P| + K)$ time [22]. We

⁵ Define a dummy node as the parent of the root node, marked with all the documents.

shall propose structures with the query cost independent of $|P|$. This section achieves a complexity of $O(\log N + K)$ by employing a GST node-numbering scheme based on the centroid path decomposition idea of Sleator and Tarjan [27]. This complexity is $O(K)$ and thus is optimal when $K \geq \log N$. The solution for the case $K < \log N$ is left to the next section.

3.1 Centroid Path Decomposition-Based Traversal

Define the *weight* of a node u in a rooted tree \mathcal{T} to be the number of nodes in the subtree rooted at u . For each internal node u , define its *successor* as its child with the maximum weight (break ties arbitrarily). The *centroid path decomposition* of tree \mathcal{T} is the spanning subgraph of \mathcal{T} whose edge set is all the predecessor-successor edges [27]. The name comes from the fact that the resulting graph consists of only disjoint paths, called *centroid paths*. A crucial property of this decomposition is that every path in \mathcal{T} can intersect only $O(\log N)$ centroid paths.

Consider the following tree-traversal algorithm defined in a recursive manner, starting at the root.

The traversal of the subtree rooted at a node u is done by first visiting node u , then recursively traversing the subtrees of the children of u . The children are ordered in such a way that the successor of u is traversed the latest.

Define the *centroid rank*, denoted as $c\text{-rank}(u)$, of a node u to be integer i if u is the i -th visited node in the aforementioned traversal algorithm. Let $c\text{-path}(u)$ be the centroid path a node u belongs to; and $c\text{-depth}(u)$ be *centroid depth*, namely, the number of the centroid paths intersected by the root-to- u path in \mathcal{T} .

We can identify the following properties.

Property 1. For each node $u \in \mathcal{T}$, the centroid ranks of all the nodes in its subtree form a range $[c\text{-rank}(u), c\text{-rank}(u) + |subtree(u)|]$.

Property 2. Let u to be a proper descendant of a node v on the same centroid path. Each descendant of u have a larger centroid rank than all descendants of v , excluding those in the subtree of u . Formally, $u' \in subtree(u)$ and $v' \in subtree(v) \setminus subtree(u)$ implies $c\text{-rank}(u') > c\text{-rank}(v')$.

3.2 The Search Structures

The reduction of [15] can be adopted to work with the numbering scheme from the centroid path decomposition-based traversal algorithm, instead of pre-order traversal. Perform the aforementioned traversal algorithm on GST. Then, a link $(origin, target, doc, score)$ qualifies a query pattern P if and only if (i) $c\text{-rank}(origin) \in [c\text{-rank}(u_P), c\text{-rank}(u_P) + |subtree(u_P)|]$, (ii) $c\text{-rank}(target) < c\text{-rank}(u_P)$, and (iii) its score is among the top K of all the links that satisfy (i) and (ii). We say a link is a *candidate* of P if it satisfies the first two requirements. Next, we categorize all the candidates of P into two types.

- *Interpath links*: those links satisfying $c\text{-depth}(target) < c\text{-depth}(u_P)$.
- *Copath links*: those links with $c\text{-depth}(target) = c\text{-depth}(u_P)$ and $c\text{-rank}(target) < c\text{-rank}(u_P)$.

In the part of this section, we shall figure out the top- K results for the two types individually. The combination of the two result sets can be achieved by a single merge in $O(K)$ time.

3.2.1 Processing Interpath Links We shall decompose the query into $O(\log N)$ subqueries. Each can be served by answering an *online sorted range reporting* query. In the online sorted range reporting problem, an array A is indexed so that given a query (i, j) , the entries in the subarray $A[i..j]$ can be reported in sorted order one by one until the user terminates the reporting. Brodal et. al. [8] proposed a linear-space structure that achieves $O(1)$ cost per entry. The top- K results among interpath links thus can be obtained by an $O(\log N)$ -way merge. Since the number of elements in the heap for merging is $O(\log N)$, an atomic heap [13] can do each heap operation in $O(1)$ time, leading to an overall $O(\log N + K)$ retrieval time.

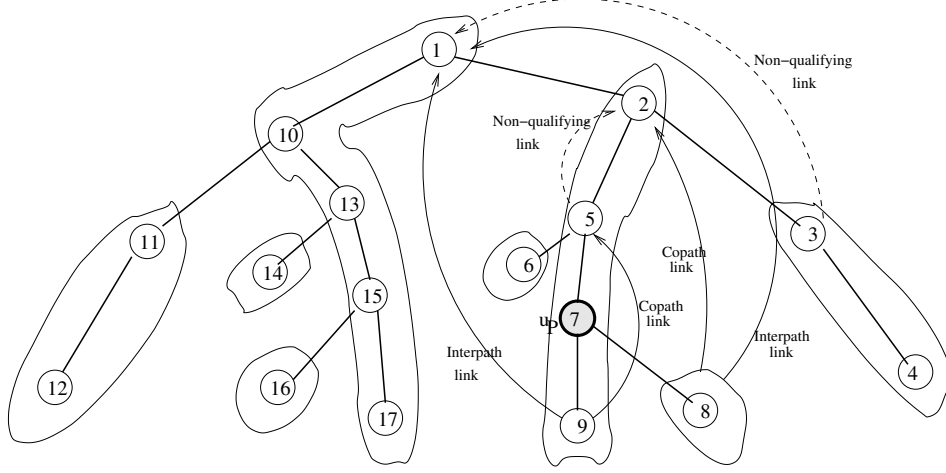


Fig. 1. Example of centroid paths, interpath and copath links.

It remains to describe how to obtain and serve the subqueries. Observe that $c\text{-depth}(\cdot)$ can have only $O(\log N)$ distinct values, meaning that we can afford to enumerate all possible values of $c\text{-depth}(\text{target})$ that satisfy $c\text{-depth}(\text{target}) < c\text{-depth}(u_P)$, one for each subquery. For each depth $\delta = 1, 2, \dots, O(\log N)$, define A_δ to be the array of all the links whose $c\text{-depth}(\text{target}) = \delta$, ordered by their $c\text{-rank}(\text{origin})$. We also keep a multiset B_δ which consists of all the $c\text{-rank}(\text{origin})$ of the links in A_δ , to support the conversion from the query range $[c\text{-rank}(u_P), c\text{-rank}(u_P) + |\text{subtree}(u_P)|)$ to the subrange $[i_\delta, j_\delta]$ on A_δ . Namely, $A_\delta[i_\delta..j_\delta]$ is exactly the set of the links in A_δ whose $c\text{-rank}(\text{origin})$ fall in the query range. Then, the subquery at level δ can be obtained by computing i_δ (resp., j_δ) as the rank of $c\text{-rank}(u_P)$ (resp., $c\text{-rank}(u_P) + |\text{subtree}(u_P)| - 1$) in B_δ . It can finally be served by an online sorted range reporting query (i_δ, j_δ) on the array A_δ .

Analysis. We can store each weighted array A_δ with the 1d top- K range reporting structure of [8], and multiset B_δ with the succinct dictionary of [26]. The decomposition into the $O(\log N)$ subqueries can be achieved in overall $O(\log N)$ time since the computation of the ranks in B_δ costs $O(1)$ time each. The retrieval time, as mentioned before, is $O(\log N + K)$. Therefore, the query cost of this part is $O(\log N + K)$. To analyze the space consumption, all the structures of A_δ take $\sum_\delta O(|A_\delta|) = O(N)$ words. As each B_δ contains at most $2N - 1$ numbers in the domain $\{1, \dots, 2N - 1\}$, its dictionary consumes $O(N)$ bits, meaning that all dictionaries consume $O(N \log N)$ bits, or $O(N)$ words. Therefore, our structure occupies $O(N)$ -word space.

3.2.2 Processing Copath Links A candidate copath link $(\text{origin}, \text{target}, \text{doc}, \text{score})$ must satisfy $c\text{-rank}(\text{origin}) \geq c\text{-rank}(u_P)$. Otherwise, as target is a proper ancestor of u_P on the same centroid path, by Property 2, origin cannot be in the subtree of u_P . On the other hand, if $c\text{-rank}(\text{target}) < c\text{-rank}(u_P) \leq c\text{-rank}(\text{origin})$, the link is a copath candidate. Therefore, the retrieval of the copath links can thus be converted into a top- K variant of the traditional 1d *interval stabbing* query [6]. Specifically, for each centroid path π , let A_π be the set of the links whose targets are on π . For each link $(\text{origin}, \text{target}, \text{doc}, \text{score}) \in A_\pi$, construct a weighted interval with (i) the left endpoint $c\text{-rank}(\text{target}) + 1$, (ii) the right endpoint $c\text{-rank}(\text{origin})$ and (iii) the weight score . Indexing the intervals properly, the query over the copath links can be served by first identifying the centroid path $\pi = c\text{-path}(u_P)$, then retrieving the top- K intervals from A_π that are stabbed by $c\text{-rank}(u_P)$.

It remains to index the weight intervals of a set A_π . Consider a sweeping line that continuously moves from $-\infty$ to $+\infty$, on which a single-linked list is maintained to keep track of all the intervals that currently intersect the sweeping line. The intervals in the linked list are sorted in descending order of their weights. As the sweeping line encounters the left (resp., right) endpoint of an interval,

it is inserted into (resp., deleted from) the linked list. For any stabbing query $c\text{-rank}(u_P)$, there must be a moment at which the first K elements of the linked list are just the answer. To support query answering on all the snapshots, the linked list can be implemented with the persistent linked list [12]. This structure guarantees that at any snapshot, once the list head has been identified, the linked list can be traversed in $O(1)$ time per element. Therefore, the top- K intervals can be retrieved by first finding the list head of the correct snapshot, which is a predecessor search; then traversing the linked list at the snapshot.

Analysis. The persistent linked list of the intervals for each A_π takes $O(|A_\pi|)$ words, implying that the overall space consumption is $\sum_\pi O(|A_\pi|) = O(N)$ since no link appears in two different sets A_π . It takes $O(\log \log |A_\pi|) = O(\log \log N)$ time to identify the list head in the persistent structure, and $O(K)$ time to report the K intervals. Therefore, the overall query cost is $O(\log \log N + K)$.

Lemma 3 *There exists a linear space data structure taking $O(\log N + K)$ time for top- K document retrieval queries once the locus of pattern is known. For $K \geq \log N$, this takes $O(K)$ time. \square*

4 Linear Space, Optimal $O(K)$ Retrieval Time Data Structure

The data structure described in the previous section is optimal for $K \geq \log N$. In this section, we provide another linear space structure, which is optimal for any $K < \log N$, thus capturing all cases.

Marked nodes and Prime nodes in a tree: Given a tree \mathcal{T} (of no single child node) of n leaves, we identify certain nodes in \mathcal{T} as marked nodes with respect to on a parameter g called *grouping factor*. The procedure starts by combining every g consecutive leaves (from left to right) together as a group, and mark the lowest common ancestor (LCA) of first and last leaf in each group. Further we mark the LCA of all pairs of marked nodes recursively. Additionally, we ensure that the root is always marked. At the end of this procedure, the number of marked nodes in \mathcal{T} will be $O(n/g)$ [15]. Every child of a marked node is called a *prime* node. For any marked node u^* , there is a *unique prime ancestor* node u' . In case u^* 's parent is marked then $u' = u^*$. For every prime node u' , the corresponding marked descendant u^* (if it exists) is unique. If u' is marked then the descendant u^* is same as u' .

4.1 The structure

Using the above scheme, we perform the marking of nodes in GST, with *grouping factor* $g = \log N$. Let u' be a prime node and let u^* (if it exists) be the unique marked descendant of u' . Then, all the links originating from the subtree of u' are categorized into the following.

1. *fringe-links*: The links originating from the subtree of u' , but not from the subtree of u^* .
2. *near-links*: The links originating from the subtree of u^* whose target is within the subtree of u' .
3. *far-link*: The link originating from the subtree of u^* whose target is a proper ancestor of u' .
4. *small-link*: The links with both origin and target within the subtree of u^* .

Lemma 4 *The number of fringe-links and the number of near-links of any prime node u' is $O(g)$.*

Proof. There are at most $2g$ leaves in $\text{subtree}(u') \setminus \text{subtree}(u^*)$. Only one link for each document comes out of the $\text{subtree}(u^*)$. Therefore, the number of *fringe-links* can be bounded by $4g$. For every document d whose link originates from $\text{subtree}(u^*)$ going out of it, it ends up as a *near-link* if and only if d exists at one of the leaves of $\text{subtree}(u') \setminus \text{subtree}(u^*)$. Thus, this can be bounded by $4g$ too. In the case that u^* does not exist for u' only fringe-links exist and since the subtree size of u' is $O(g)$ there can be no more than $O(g)$ of these links. \square

Consider the following set, consisting of $O(g)$ links with respect to u' : all *fringe-links*, *near-links* and g highest scored *far-links*. For any node u , whose closest prime ancestor (including itself) is u' , the above mentioned set is called *candidate links* of u . We maintain this *candidate links* at u' . From each u , we maintain the pointer to its closest prime ancestor where the list of *candidate links* is stored.

Lemma 5 *The candidate links of any node u contains top- g highest scored links among those with origin below the subtree of u and target above the subtree of u .*

Proof. Let u' be the closest prime ancestor of u . If no marked descendant of u' exist, then all the links are stored as candidate links. Otherwise, *small-links* can not ever be candidates as they never cross u . Now, if u lies on the path from u' to u^* then all *far-links* will satisfy both origin and target conditions. Else, *far-link* do not qualify. Hence, any link which is not among top- g (highest scored) of these far-links, can never be the candidate. \square

We maintain all these candidate links in the sorted order of score (as a list called *candidate list*), and maintain a pointer to it from all those node u whose top- g links belongs to this collection. Note that the candidate list of many nodes can be the same, and it consists of $O(g)$ links. To filter out the top- K links corresponding any node u , we maintain additional structures.

Let B_u be a bit vector of length $O(g)$ associated with node u , such that $B_u[i] = 1$ if and only if the i th highest scored link in the *candidate list* of u (maintained at u') is a *valid* link for u . A link is said to be *valid* with respect to a node u if its origin is in the subtree of u and target above u . Since this bit-vector is of length $O(g) = O(\log N)$, we can easily implement rank/select structure on it (using tables) which can give answers in $O(1)$ time.

4.2 Query Answering

In order to answer the top- K query (for $K < g$) corresponding to a locus node u_P , we just retrieve those $\text{select}(B_{u_P}, i)$ th highest scored link in the candidate list of u_P (stored at its prime ancestor u'_P) for $i = 1, 2, 3, \dots, K$, where $\text{select}(B_{u_P}, i)$ returns the position of i th 1 in the bit vector B_u . These select queries give the positions corresponding to the location of a *valid* links for u_P . Since the links are sorted in the score order, we get the top- K answers in sorted order.

Space-Time Analysis: The total space for maintaining $O(g)$ words *candidate links* at every $O(N/g)$ marked nodes is $O(N)$. By choosing $g = \log N$, the total space of bit vectors associated with all nodes can be bounded by $O(N \log N)$ bits, and the number of pointers is also $O(N)$. The query time is $O(K)$ as each select query takes only constant time.

Lemma 6 *There exists a data structure taking $O(N)$ space which can answer top- K document retrieval queries in optimal $O(K)$ time, for any $K < \log N$.* \square

Combining Lemma 6 with Lemma 3 we get our main theorem.

Theorem 1 *There exists a data structure taking $O(N)$ space which can answer top- K document retrieval queries in optimal $O(K)$ time, once the locus node of the pattern is known.* \square

5 External Memory Data Structures

It is known that no linear-space external memory structure can answer the (even the simpler) 1d top- K range reporting query in $O(\log^{O(1)} N + K/B)$ I/Os if the output order must be ensured. We thus turn our attention to solving the unordered variant of the top- K document retrieval problem in the external memory and cache-oblivious models. Namely, the K results can be returned in an arbitrary order.

As the reduction from top- K document retrieval to top- K hinged range reporting still work in external memory (refer to Section 2), this section further converts the top- K requirement into a one-sided score constraint, called a *threshold*, then solving the resulting problem with a divide-and-conquer idea. The problem can be formally stated as follows.

Problem 1. Index a set S of 3d points⁶, so that a query (a, b, τ) returns all points $(x, y, z) \in S$ satisfying $y < a \leq x \leq b$ and $z \leq \tau$.

⁶ The third dimension comes from the scores.

5.1 Converting Top- K to Threshold via the Logarithmic Sketch

By setting the grouping factor g again to $\log N$ and computing marked and prime nodes as in Section 4, a similar query answering scheme can be adopted. Specifically, given the locus u_P of a query P whose lowest prime ancestor is u' , we process *fringe*-, *near*- and *far-links* with respect to u' individually. The *fringe*- and *near*- links can be handled by simply scanning and checking all of them against the query constraints, and finally returning the K links with the highest scores (in case less than K links qualify, return all of them). This takes only $O(g)/B = O(\frac{\log N}{B}) = O(\log_B N)$ I/Os.

The processing of *far-links*, however, is more sophisticated. As the number of *far-links* with respect to u' can be much larger than $\log N$, we cannot afford storing all of them. Instead, we keep a logarithmic sketch of the *far-links* of u' . Namely, the sketch consists of the scores of the *far-links* that rank the first, second, fourth, eighth, and so on. The top- K results among the *far-links* can be retrieved by the following steps. First, find the score τ that ranks the $2^{\lceil \log K \rceil}$ -th, which has been stored in the logarithmic sketch on u' . Then, with the found τ and the tree range of u_P , issue a query of Problem 1 over all links. This step will return, instead of top- K , top- $O(K)$ links.

The combination of the two result sets can be done by a K -selection algorithm over the $K + O(K) = O(K)$ links. Since a logarithmic sketch takes $O(\log N)$ -word space, and the number of prime nodes is $O(N/g) = O(N/\log N)$, the overall space consumption, excluding the structure of Problem 1, is linear. And the query cost, excluding the subquery of Problem 1, incurs $O(\log_B N + K/B)$ I/Os. Therefore, to achieved the claimed bound, it remains to solve Problem 1 with a linear-space structure that answers a query in $O(\log_B N + \log \log B + K/B)$ I/Os.

5.2 Near I/O-Optimal, Linear Space Data Structure

5.2.1 Small-Grid Structure Given an additional restriction to Problem 1 that every point $(x, y, z) \in S$ satisfies $x, y \in \{1, 2, \dots, U\}$, this subsection proposes a data structure that takes $O(|S| + U^2 B)$ -word space, and answers a query in $O(\log_B |S| + K/B)$ I/Os. Such a structure will require the following toolkit.

Lemma 1 (Persistent Sorted List). *Given an update sequence (of insertions and deletions) on a total-ordered set, there exists a linear-space data structure that can retrieve the $Z > 0$ minimum elements at **any version** in $O(1 + Z/B)$ I/Os. Here, a version is the content of the set after an update.*

Proof. (sketch) Consider a block-based linked list implementation. Each node in the linked list is a block, which stores (i) $B/2$ to B elements, and (ii) a pointer to its next node in the linked list. Then, insertions and deletions can be handled easily: if an insertion to some node makes it contain more than B elements, split the node into two nodes; if a deletion makes it contain less than $B/2$ node, merge it with its predecessor or successor (both work), then split the new node if necessary. This structure can be easily made persistent with standard persistent techniques [12,4].

Now, for each integer $x_0 \in \{1, \dots, U\}$, apply Lemma 1 on all the points whose x-coordinates are x_0 : regard each point (x_0, y, z) as an element with sorting key z , inserted at time y and never deleted. Therefore, we have obtained U persistent sorted lists. Based on them, a query (a, b, τ) can be answered in $O(U + K/B)$ I/Os: for every persistent sorted list whose points have x-coordinates in $[a, b]$, first pick the version on time a , then report all the elements whose keys are below τ .

It remains to replace the term U with $\log_B |S|$ in the query cost. To achieve this, we will adopt the idea of the external memory priority search tree [3] to build some “top” structures. For each $x_0, y_0 \in \{1, \dots, U\}$, let

$$S[x_0, \leq y_0] = \{(x, y, z) \in S \mid x = x_0, y \leq y_0\}$$

and

$$S[\leq y_0] = \bigcup_{x \geq y_0} \text{top-}B(S[x, \leq y_0]),$$

where $\text{top-}Z(Q)$ represents the Z points in Q with the minimum z-coordinates. In other words, $S[\leq y_0]$ picks the top- B elements from each relevant persistent sorted list at version y_0 . Then, we can answer

the query (a, b, τ) against $S[\leq a]$ first. If less than B points are reported for a specific x_0 , we know that all resulting elements in that persistent sorted list have been reported; thus, we can skip it. If, on the other hand, at least B points are reported, we can afford querying the persistent sorted list since we will report at least B points from it. We have thus eliminated the term U , but introduced a new term due to the search of $S[\leq a]$.

To efficiently retrieve the points in $S[\leq a]$ that qualifies, we know that each point $(x, y, z) \in S[\leq a]$ already satisfies $y < a \leq x$. So we only need to issue a quadrant query to find all the points satisfying $x \leq b$ and $z \leq \tau$. We use the external memory priority search tree, whose query cost is $O(\log_B |S[\leq a]| + Z/B)$ if Z points qualify. Therefore, the overall query cost is $O(\log_B |S| + K/B)$.

To analyze the space consumption, observe that no two persistent sorted lists share the same point. Thus, the persistent sorted lists consume $O(|S|)$ space. Then, for each $y_0 \in \{1, \dots, U\}$, an external memory priority search tree is built, which takes $O(UB)$ words. Overall, the structure occupies $O(|S| + U^2B)$ words.

5.2.2 A General Structure This subsection bootstraps the data structure proposed in the previous subsection to solve Problem 1 in the general setting. Consider the following divide-and-conquer scheme. Let $N' = N^{2/3}B^{1/3}$. Partition the point set S into S_1, S_2, \dots, S_L according to the x-coordinates, such that (i) each set S_i contains $N'/2$ to N' points for $i = 1, \dots, L$, and (ii) all points in S_i have less x-coordinates than any point in S_{i+1} for $i = 1, \dots, L-1$. Denote by δ_i ($i \in [1, L]$) the minimum x-coordinate in S_i .⁷ Therefore, $\delta_2, \delta_3, \dots, \delta_L$ partitions the 3d space into L vertical “slabs”. We further partition the slabs according to planes $y = \delta_2, y = \delta_3, \dots, y = \delta_L$. As a result, the whole space is partitioned into $O(L^2)$ cells. Figure 2(a) illustrates the xy-projection of the partition. Now, given a

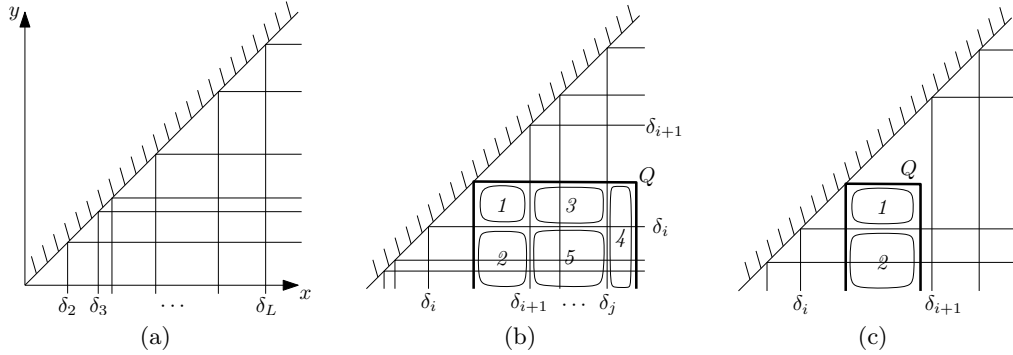


Fig. 2. Divide-and-conquer scheme in EM. Only x- and y-coordinates are illustrated.

query $Q = (a, b, \tau)$, we decompose it into five subqueries Q_1, Q_2, \dots, Q_5 as illustrated in Figure 2(b). Formally, let i, j be integers that satisfy $\delta_i \leq a < \delta_{i+1}$ and $\delta_j \leq b < \delta_{j+1}$. Then, the subqueries are defined as follows.

- Q_1 is the part of Q whose xy-projection falls in $[\delta_i, \delta_{i+1}) \times [\delta_i, \delta_{i+1})$.
- Q_2 is the part of Q whose xy-projection falls in $[\delta_i, \delta_{i+1}) \times [-\infty, \delta_i)$.
- Q_3 is the part of Q whose xy-projection falls in $[\delta_{i+1}, \delta_j) \times [\delta_i, \delta_{i+1})$.
- Q_4 is the part of Q whose xy-projection falls in $[\delta_j, \delta_{j+1}) \times [-\infty, \delta_{i+1})$.
- Q_5 is the part of Q whose xy-projection falls in $[\delta_{i+1}, \delta_j) \times [-\infty, \delta_i)$.

It is also possible that $i = j$. This degenerated case is shown in Figure 2(c), in which only Q_1 and Q_2 may contain result points. Either way, in general, subqueries Q_2, Q_3 and Q_4 can be answered with either 2d 3-sided range queries [3] or 3d dominance search [1] (recall that the z-coordinates are not illustrated here). Therefore, they can all be answered in $O(\log_B N + Z/B)$ I/Os if Z points are

⁷ For notational convenience, specially define $\delta_{L+1} = +\infty$.

retrieved. Furthermore, Q_1 is a subproblem with exactly the same definition as Problem 1, but with problem size N' instead of N .

It remains to consider Q_5 . A crucial observation is that all its three boundaries (about x - and y -coordinates) are on the partition planes. Furthermore, the left boundary (i.e., plane $x = \delta_{i+1}$) is always the successor of the top boundary (i.e., plane $y = \delta_i$). Therefore, we can reassign the x - and y -coordinates of the points such that Q_5 becomes a small-grid problem which has been solved in Section 5.2.1. The reassignment works as follows. Given a point (x, y, z) , reassign x to integer k if $\delta_k \leq x < \delta_{k+1}$; reassign y to integer $k + 1$ if $\delta_k \leq y < \delta_{k+1}$. Therefore, Q_5 can also be answered in $O(\log_B N + Z/B)$ I/Os. As no point is reported twice, by ignoring the necessary $O(K/B)$ -I/O term to retrieve $O(K)$ documents, the remaining part of the query cost is given by the following equation.

$$\mathcal{T}(N) = \mathcal{T}(N^{2/3}B^{1/3}) + O(1 + \log_B N). \quad (1)$$

The recursion terminates with $\mathcal{T}(N) = O(1)$ for $N = O(B)$.

From now on, we use N_0 to denote the initial problem size, and N to denote the current problem size (due to the recursion). To solve (1),

$$\begin{aligned} \mathcal{T}(N_0) &= \mathcal{T}((N_0/B) \cdot B) \\ &= \mathcal{T}((N_0/B)^{2/3} \cdot B) + O(1 + \log_B(N_0/B)) \\ &= \mathcal{T}((N_0/B)^{(2/3)^2} \cdot B) + O(1 + \log_B(N_0/B)) + O(1 + \log_B(N_0/B)^{2/3}) \\ &= O\left(\sum_{k=0,1,\dots} 1 + (2/3)^k \log_B(N_0/B)\right) \\ &= O(\log \log N_0 + \log_B(N_0/B)), \end{aligned}$$

where the term $\log \log N_0$ comes from the fact that the recursion level is upper bounded by $O(\log \log N_0)$. Let $\mu = \log_B N_0$. Then,

$$\log \log N_0 + \log_B N = \log \log(B^\mu) + \mu = \log \mu + \log \log B + \mu = O(\log \log B + \mu).$$

We have thus proved the desired query bound.

Space Consumption In the divide and conquer, observe that the point sets used in answering subqueries Q_2, \dots, Q_5 are totally disjoint with the subproblem of Q_1 . In other words, there is a “non-duplication” property here: once a point appears in a structure to answer any subqueries Q_2, \dots, Q_5 , it does not appear in any recursive subproblem. Therefore, by traversing the recursion tree and counting the space consumptions of the structures for these subqueries, this part contributes a cost of

$$\sum_{\text{node } v} (O(N_v) + O(N_v + U_v^2 B)),$$

where N_v is the number of points used in the recursion node v , and U_v is the parameter L at v . The summation of $O(N_v)$ is $O(N_0)$. We now compute the summation of $O(U_v^2 B)$. As $N' = N_v^{2/3} B^{1/3}$ at node v , parameter $L \geq 2N_v/N' \leq 2(N_v/B)^{1/3}$. Therefore, at the root level, we have one term with $U_v^2 = 4(N_0/B)^{2/3}$; at one level down, we have at most $2N_0/N_1$ terms with $U_v^2 \leq 4(N_1/B)^{2/3}$ where $N_1/B = (N_0/B)^{2/3}$; in general, at level l (level 0 is the root level), we have at most $2N_0/N_l$ terms with $U_v^2 \leq 4(N_l/B)^{2/3}$ where $N_l/B = (N_0/B)^{(2/3)^l}$. Therefore, the summation of U_v^2 at level l is at most

$$(2N_0/N_l) \cdot 4(N_l/B)^{2/3} = 8(N_0/B)(B/N_l)^{2/3} = 8(N_0/B)(B/N_0)^{(2/3)^{l+1}}.$$

Notice that the term $(B/N_0)^{(2/3)^{l+1}}$ degenerates exponentially as l increases. Therefore, their summations over l is dominated by the first term $8(N_0/B)^{1/3}$, meaning that $\sum_v O(U_v^2 B) = O((N_0/B)^{1/3} B) = O(N_0)$.

Combining with string B-tree [7], we get the following result.

Theorem 2 *There exists a linear space data structure taking $O(N)$ words of space, which answers top- K (unsorted) document retrieval queries in $O(|P|/B + \log_B N + \log \log B + K/B)$ I/Os. \square*

5.3 I/O-Optimal, Almost Linear Space Data Structure

Clearly the data structure in sec 5 is optimal for $K = \Omega(B \log \log B)$. The case when $K < B \log \log B$ can be handled separately based on the result in lemma 5 as follows: we maintain the *candidate links* as a list in the sorted order of score. Then top- K documents ($K < g$, the grouping factor) can be retrieved in $O(1 + g/B)$ I/O's in the *sorted order of score* by scanning all candidate links and reporting those which satisfy the origin-target conditions with respect to u_P . Note that the query time is optimal when $K = \Theta(g)$, and the space requirement is $O(N)$ words. Therefore by maintaining the above described linear space structure for $g = B, 2B, 4B, \dots, B \log \log B$, the top- K query corresponding to any $K < B \log \log B$ can be answered optimally in $O(1 + K/B)$ -I/O's by querying on the structure corresponding to $K \leq g < 2K$.

Theorem 3 *There exists an almost-linear space data structure of $O(N \log \log \log B)$ -word space, which answers top- K (unsorted) document retrieval queries in optimal- $O(|P|/B + \log_B N + K/B)$ I/Os. \square*

6 Cache-Oblivious Data Structures

Our cache-oblivious results are derived using the internal memory framework of sec 3 and 4, where all internal memory data structures are replaced by the best known corresponding cache-oblivious counterparts, and we get the following result.

Theorem 4 *There exists cache-oblivious data structures for the top- K (unsorted) document retrieval problem with the following space-I/O trade-off's.*

- $O(N\sqrt{\log N} \log \log N)$ -word space and $O(\log N \log_B N + K/B)$ -I/Os.
- $O(N\sqrt{\log N} \log^2 \log N)$ -word space and $O(\log \log N \log_B N + K/B)$ -I/Os.

Proof. (sketch) The top- K document retrieval problem can be converted to its threshold version via the logarithmic sketch (sec 5.1). Using the framework in sec 3, the original problem can be decomposed into $O(\log N)$ three-sided queries and a 3-d dominance query (general case of stabbing intervals with score). By substituting an $O(N\sqrt{\log N} \log \log N)$ -word space and $O(\log_B N + K/B)$ -I/Os data structure [2] for these sub-problems, the first result can be obtained. As the input range is the same for all $O(\log n)$ three sided queries, the structures can be combined⁸ and reduce the number of three sided queries to $O(\log \log N)$ with $O(\log \log N)$ blowup in space. This leads to the second result. \square

Theorem 5 *There exists a linear space data structure of $O(N)$ -word space, which answers document listing queries in $O(\log N + ndoc/B)$ I/Os, where $ndoc$ is the number of documents containing P .*

Proof. (sketch) The document listing problem (i.e., without score constraint) can be reduced to scanning and reporting of elements from $O(\log N)$ lists, hence can be served in $O(\log N + ndoc/B)$ I/Os. \square

⁸ Consider a balanced binary tree of $\log N$ leaves, thus $O(\log \log N)$ height. The three-sided range reporting over the links corresponding to $c\text{-depth}(\text{target}) = i$ is stored at leaf i . An internal node u maintain a combined three-sided range reporting structure of all those links with $c\text{-depth}(\text{target}) = j$, given the j th leftmost leaf is in the sub-tree of u . Here each link is a part of $O(\log \log N)$ structures, hence the space will blowup by an $O(\log \log N)$ factor. However the the number of structures to be searched is reduced to $O(\log \log N)$.

References

1. Peyman Afshani. On dominance reporting in 3d. In *ESA*, pages 41–51, 2008.
2. Peyman Afshani, Chris H. Hamilton, and Norbert Zeh. Cache-oblivious range reporting with optimal queries requires superlinear space. *Discrete & Computational Geometry*, 45(4):824–850, 2011.
3. Lars Arge, Vasilis Samoladas, and Jeffrey Scott Vitter. On two-dimensional indexability and optimal range search indexing. In *PODS*, pages 346–357, 1999.
4. Bruno Becker, Stephan Gschwind, Thomas Ohler, Bernhard Seeger, and Peter Widmayer. An asymptotically optimal multiversion b-tree. *VLDB J.*, 5(4):264–275, 1996.
5. Djamel Belazzougui and Gonzalo Navarro. Improved compressed indexes for full-text document retrieval. In *SPIRE*, pages 386–397, 2011.
6. Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag TELOS, Santa Clara, CA, USA, 3rd ed. edition, 2008.
7. Gerth Stølting Brodal and Rolf Fagerberg. Cache-oblivious string dictionaries. In *SODA*, pages 581–590, 2006.
8. Gerth Stølting Brodal, Rolf Fagerberg, Mark Greve, and Alejandro López-Ortiz. Online sorted range reporting. In *ISAAC*, pages 173–182, 2009.
9. Bernard Chazelle. Lower bounds for orthogonal range searching: I. the reporting case. *J. ACM*, 37(2):200–212, 1990.
10. Yu-Feng Chien, Wing-Kai Hon, Rahul Shah, and Jeffrey Scott Vitter. Geometric burrows-wheeler transform: Linking range searching and text indexing. In *DCC*, pages 252–261, 2008.
11. J. Shane Culpepper, Gonzalo Navarro, Simon J. Puglisi, and Andrew Turpin. Top- k ranked document search in general text databases. In *ESA (2)*, pages 194–205, 2010.
12. James R. Driscoll, Neil Sarnak, Daniel Dominic Sleator, and Robert Endre Tarjan. Making data structures persistent. *J. Comput. Syst. Sci.*, 38(1):86–124, 1989.
13. Michael L. Fredman and Dan E. Willard. Trans-dichotomous algorithms for minimum spanning trees and shortest paths. *J. Comput. Syst. Sci.*, 48(3):533–551, 1994.
14. Wing-Kai Hon, Rahul Shah, and Sharma V. Thankachan. Towards an optimal space-and-query-time index for top- k document retrieval. In *CPM*, pages 173–184, 2012.
15. Wing-Kai Hon, Rahul Shah, and Jeffrey Scott Vitter. Space-efficient framework for top- k string retrieval problems. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '09, pages 713–722, Washington, DC, USA, 2009. IEEE Computer Society.
16. Wing-Kai Hon, Rahul Shah, and Jeffrey Scott Vitter. Compression, indexing, and retrieval for massive string data. In *CPM*, pages 260–274, 2010.
17. Marek Karpinski and Yakov Nekrich. Top- k color queries for document retrieval. In *SODA*, pages 401–411, 2011.
18. Gregory Kucherov, Yakov Nekrich, and Tatiana A. Starikovskaya. Cross-document pattern matching. In *CPM*, pages 196–207, 2012.
19. M. Oguzhan Külekci, Jeffrey Scott Vitter, and Bojian Xu. Efficient maximal repeat finding using the burrows-wheeler transform and wavelet tree. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 9(2):421–429, 2012.
20. Yossi Matias, S. Muthukrishnan, Süleyman Cenk Sahinalp, and Jacob Ziv. Augmenting suffix trees, with applications. In *Proceedings of the 6th Annual European Symposium on Algorithms*, ESA '98, pages 67–78, London, UK, UK, 1998. Springer-Verlag.
21. S. Muthukrishnan. Efficient algorithms for document retrieval problems. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 657–666, 2002.
22. Gonzalo Navarro and Yakov Nekrich. Top- k document retrieval in optimal time and linear space. In *SODA*, pages 1066–1077, 2012.
23. Gonzalo Navarro and Simon J. Puglisi. Dual-sorted inverted lists. In *SPIRE*, pages 309–321, 2010.
24. Manish Patil, Sharma V. Thankachan, Rahul Shah, Wing-Kai Hon, Jeffrey Scott Vitter, and Sabrina Chandrasekaran. Inverted indexes for phrases and strings. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 555–564, New York, NY, USA, 2011. ACM.
25. Manish Patil, Sharma V. Thankachan, Rahul Shah, Wing-Kai Hon, Jeffrey Scott Vitter, and Sabrina Chandrasekaran. Inverted indexes for phrases and strings. In *SIGIR*, pages 555–564, 2011.
26. Rajeev Raman, Venkatesh Raman, and Srinivasa Rao Satti. Succinct indexable dictionaries with applications to encoding k -ary trees, prefix sums and multisets. *ACM Transactions on Algorithms*, 3(4), 2007.
27. Daniel Dominic Sleator and Robert Endre Tarjan. A data structure for dynamic trees. *J. Comput. Syst. Sci.*, 26(3):362–391, 1983.
28. Niko Välimäki, Susana Ladra, and Veli Mäkinen. Approximate all-pairs suffix/prefix overlaps. In *CPM*, pages 76–87, 2010.
29. Niko Välimäki and Veli Mäkinen. Space-efficient algorithms for document retrieval. In *CPM*, pages 205–215, 2007.
30. Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM Comput. Surv.*, 38(2), July 2006.